SOLUTION BRIEF

SUPERMICRO AND F5 TOGETHER DELIVER COMPREHENSIVE SECURITY, PERFORMANCE, AND MANAGEMENT SOLUTIONS

Table of Contents

Executive Summary

Combine F5's advanced Security, Scalability, and Performance capabilities with Supermicro's high-performance, customizable hardware solutions to deliver unparalleled application security and provide enterprises a robust, scalable, and integrated platform to power their digital transformation initiatives while ensuring optimal application performance and security.

Key Benefits

As NVIDIA customers expand their AI capabilities, they increasingly rely on massive clusters of NVIDIA's GPUs to build and train sophisticated AI models. These models are not only growing in complexity but also in the time required for training and inferencing due to:

- Scaling Issues: Network congestion of north-south traffic to and inside the AI clusters, coupled with the immense computational demands requiring trillions of operations per second, significantly slows down applications and processes.

- Security Concerns: Many AI models process highly sensitive information, necessitating robust security measures to protect data integrity.

• Observability limits: AI clusters face challenges pinpointing performance bottlenecks affecting training or inference workloads due to a lack of observability insights, resulting in inefficiencies and resource wasted resources.

BIG-IP Next for Kubernetes

BIG-IP Next for Kubernetes deployed on NVIDIA BlueField-3 DPUs integrates F5's industry-leading network traffic management and security solutions with NVIDIA accelerated computing, providing a unified solution optimized for AI workloads. The comprehensive solution natively encompasses an edge firewall, DDoS mitigation, and intrusion prevention systems, offering robust protection against emerging threats without increasing the hardware footprint. This integration ensures AI networks remain secure and performant amidst growing vulnerability concerns. Key features include:

• Performance Optimization: F5 BIG-IP now runs on NVIDIA's high-performance Data Processing Units (DPUs), alleviating the CPU from the processing workload of networking, traffic management, and security tasks. This allows GPUs to focus more efficiently on AI tasks. During a session at NVIDIA GTC, SoftBank shared game-changing insights on how organizations can turbocharge cloud-native AI workloads with a DPU-accelerated service proxy for Kubernetes. The session featured SoftBank's calculations and performance metrics from their recent proof of concept of F5 BIG-IP Next for Kubernetes deployed on NVIDIA BlueField-3 DPUs. SoftBank achieved an 18% increase in HTTP throughput (77 Gbps), an 11x improvement in time-to-first-byte (TTFB), and a staggering 190x boost in network energy efficiency. These results highlight the transformative potential of DPU acceleration for modern cloud-native environments, driving improved throughput of tokens and enhanced user experiences during AI inferencing.

Customers deploying this solution achieve performance improvements, unlocking greater value and efficiency.

• Enhanced Security: F5 BIG-IP Next for Kubernetes deployed on NVIDIA BlueField-3 DPUs offers end-to-end encryption and advanced security capabilities, supporting zero-trust security models and protecting sensitive data. Further, it enhances supply chain security and helps with achieving PCI compliance.

• Traffic management and Load Balancing: Facilitates efficient and smart network traffic distribution and resource allocation among multiple tenants and multiple GPUs, optimizing their overall usage.

• Versatility: First deployment of F5 solutions on ARM chipsets, optimizing energy consumption.

Such a cohesive solution significantly outperforms alternatives by unifying networking, security, traffic management, and observability into a single pass. This unified approach significantly improves performance and simplifies deployment, management, and operations, eliminating the need for disparate products and considerably saving time, resources, energy, and hardware costs.

Supermicro Systems

High-Performance, Customizable Hardware:

• Tailored Solutions: Supermicro offers highly customizable hardware configurations to meet specific needs, ensuring that the integrated F5 solutions run at peak performance.

• Scalability: Scale your infrastructure effortlessly with Supermicro's wide range of server solutions, designed to grow with your business needs.

• Reliability and Support: Benefit from Supermicro's robust global support and service infrastructure, ensuring continuous operation and minimal downtime.

Ideal Supermicro systems for these use cases include:

ARS-221GL-NR:

- 2U Rackmount
- NVIDIA Dual 72-core CPUs on an NVIDIA Grace™ Superchip
- 960GB ECC LPDDR5X
- Up to 2 double-width GPUs

AS -4125GS-TNRT

- 4U Rackmount
- Dual AMD EPYC™ 9004/9005 Series Processors
- Up to 9.2TB DDR5-5200 MT/s Memory (9005 Series)
- Up to 8 double-width GPUs

ARS-111GL-NHR

- 1U Rackmount
- NVIDIA 72-core NVIDIA Grace CPU on GH200 Grace Hopper™ Superchip
- 480GB ECC LPDDR5X (+ 96GB ECC HBM3) memory
- Up to 1 double-width GPU

SYS-221H-TNR

- 2U Rackmount
- Dual 4th / 5th Gen Intel® Xeon® Scalable processors
- Up to 4TB (1DPC) DDR5-5600MT/s; Up to 8TB (2DPC) DDR5-4400MT/s memory
- Up to 4 double-width GPUs

Use Cases

- Financial Services: Secure and accelerate financial applications, ensuring compliance and protecting sensitive customer data.
- Healthcare: Enhance the performance and security of healthcare applications, safeguard patient information, and improve access to medical data.
- Telecommunications: Optimize network performance and security for telecommunications providers, supporting high traffic volumes and ensuring reliable service delivery.
- Government and Education: Provide secure, high-performance solutions for government and educational institutions, protecting sensitive information and improving application access and performance.

Conclusion:

The synergy between F5 BIG-IP Next for Kubernetes's capabilities and Supermicro's high-performance hardware solutions offers a powerful, secure, and efficient platform for enterprises looking to optimize their application delivery and security. This joint value proposition ensures that organizations can confidently navigate the complexities of digital transformation, delivering exceptional user experiences while safeguarding critical data and applications.

Further Information

www.supermicro.com

https://www.f5.com/products/big-ip/next/kubernetes-on-nvidia-bluefield-dpu

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.

Learn more at www.supermicro.com

F5

F5 is a multicloud application security and delivery company committed to bringing a better digital world to life. F5 partners with the world's largest, most advanced organizations to secure every app—on premises, in the cloud, or at the edge. F5 enables businesses to continuously stay ahead of threats while delivering exceptional, secure digital experiences for their customers. For more information, go to f5.com. (NASDAQ: FFIV)

